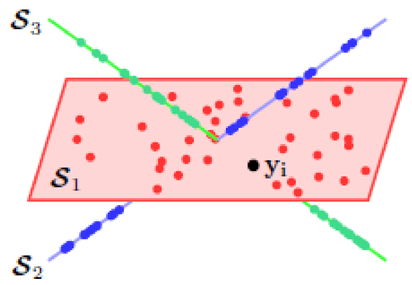
A Deterministic Analysis of Noisy Sparse Subspace Clustering for Dimensionality-reduced Data

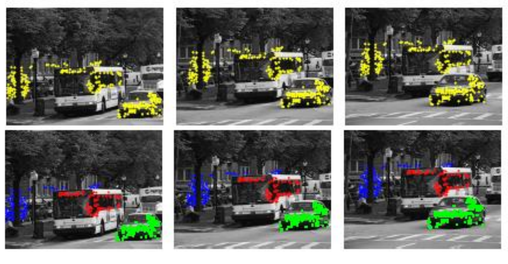
Yining Wang, Yu-Xiang Wang and **Aarti Singh**
Carnegie Mellon University, Machine Learning Department

Subspace clustering: clustering data points into union of low-dimensional subspaces



Mathematically: given $x_1, \cdots, x_N \in R^d$, find linear subspaces $S_1, \cdots, S_L$ of dimension $r \ll d$ such that each x_i approximately lies in some S_k

Applications: motion segmentation

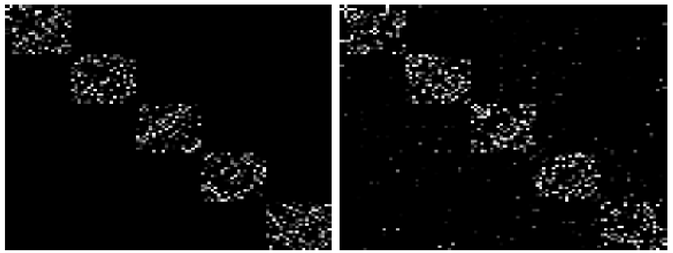


… and many more: face clustering, network hop counting, social graph mining, recommendation systems …

Sparse Subspace Clustering (SSC, Elhamifar & Vidal 2007): state-of-the-art subspace clustering algorithm based on ℓ_1 self-expression

Step 1. Instance-level ℓ_1 self-regression
$\hat{c}_i = \text{argmin}_{c \in R^{N-1}} \{\|x_i - cX_{-i}\|_2^2 + \lambda\|c\|_1\}$

Step 2. Build similarity graph $G \in R^{N \times N}$ by taking $G_{ij} = |\hat{c}_{ij}| + |\hat{c}_{ji}|$



Step 3. Spectral clustering on similarity graph G

Question: will SSC still succeed if the ambient data dimension d is reduced to $p \ll d$ by linear dimensionality reduction?

$$\widetilde{X} = \Psi X, \qquad \Psi \in R^{p \times d}$$

Motivation: computational efficiency, compressed measurement, missing data, data privacy, etc.

Method: Gaussian projection, Fast Johnson-Lindenstrauss transform (FJLT), uniform row sampling, sketching, etc.

Property: **subspace embedding property**
$\Pr[\forall x \in S, \|\Psi x\|_2^2 \in (1 \pm \epsilon)\|x\|_2^2] \geq 1 - \delta$

Deterministic analysis of Noisy Sparse Subspace Clustering under dimension reduction

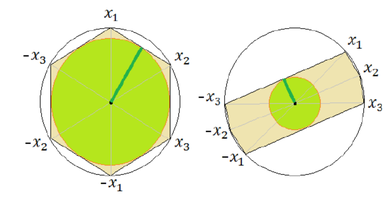
Subspace incoherence: for subspace S_ℓ define

$$\mu_\ell = \max_{x \in X \setminus X^{(\ell)}} \left\| V^{(\ell)^T} x \right\|_\infty$$

where $V^{(\ell)} = \{\text{normalize}(P_{S\_\ell}[v(x_i^{(\ell)})])\}$ and $v(x)$ is the optimal solution to dual problem

$$\max_{v \in R^d} \langle v, x \rangle + 0.5\lambda\|v\|_2^2, s.t. \|X^T v\|_\infty \leq 1$$

Inradius: ρ_ℓ characterizing inner-subspace data distribution



No false connection: $(x_i, x_j) \in E(G) \Longrightarrow x_i, x_j$ belong to the same cluster (subspace).

Main Theorem **Let η be the level of adversarial noise, ϵ be the parameter in subspace embedding property and $\Delta = \min_\ell(\rho_\ell - \mu_\ell)$ be the *geometric gap*. Then G has no false connections with high probability if**

$$\epsilon \leq \min\left\{\frac{1}{3}, \frac{\Delta}{4(2+\rho)}, \frac{\lambda}{8}\left(c_2\Delta^2 - \frac{5\eta^2}{\rho}\right) - 3\eta\right\}$$